Web browser as a tool for predicting the incidence of influenza

Przeglądarka internetowa jako narzędzie do przewidywania zapadalności na grypę

Sylwia W. Wójcik[1,A–D], Mariusz Duplaga[1,A,C,E–F], Marcin Grysztar[1,C–D], Paulina Pałka[2,B,D]

[1] Department of Health Promotion, Institute of Public Health, Faculty of Health Sciences, Jagiellonian University Medical College, Cracow, Poland
[2] Students' Scientific Circle of Health Promotion, Institute of Public Health, Faculty of Health Sciences, Jagiellonian University Medical College, Cracow, Poland

A – research concept and design; B – collection and/or assembly of data; C – data analysis and interpretation;
D – writing the article; E – critical revision of the article; F – final approval of the article

Pielęgniarstwo i Zdrowie Publiczne, ISSN 2082-9876 (print), ISSN 2451-1870 (online) *Piel Zdr Publ.* 2018;8(2):83–88

Address for correspondence
Sylwia Wójcik
E-mail: sw.wojcik@gmail.com

Funding sources
This study was supported by resources of the statutory project of the Jagiellonian University Medical College, No. K/ZDS/006112.

Conflict of interests
None declared

Received on November 22, 2017
Reviewed on January 9, 2018
Accepted on February 2, 2018

Abstract

Background. Infodemiology is focused on the analysis of web content to predict health phenomena. Google Trends (GT) is a free and publicly available service that permits analyses of searches performed with the Google web search engine. With GT it is possible to specify how often certain keywords are searched for.

Objectives. The purpose of the study was to determine the feasibility of using data on the frequency of searches with the Google search engine to predict influenza incidence.

Material and methods. Using the GT service, data on the frequency of searches for the Polish equivalents of "flu", "cold" and "fever" in the period of 2014–2016 in Poland were retrieved. Simultaneously, the epidemiological reports prepared by the National Institute of Public Health – National Institute of Hygiene (NIPH-NIH) were obtained for influenza incidence in the same period. Correlations between the variables were assessed using Spearman's rank-order correlation.

Results. A statistically significant correlation was confirmed between the average daily search coefficients (ADSC) for all 3 keywords and weekly influenza incidence according to the NIPH-NIH data. The strongest correlation was found for the ADSC of the word "cold" ($r = 0.77$; $p < 0.05$).

Conclusions. The frequency of searches implemented with the Google search engine may be used for predicting the incidence of influenza in the Polish population.

Key words: infodemiology, influenza, Internet searches

DOI
10.17219/pzp/84984

Copyright
© 2018 by Wroclaw Medical University
This is an article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Streszczenie

Wprowadzenie. Infodemiologia zajmuje się analizą treści internetowych w celu przewidywania zjawisk zdrowotnych. Google Trends (GT) to bezpłatny i publicznie dostępny serwis, który pozwala na analizę wyszukiwań w wyszukiwarce internetowej Google. Udostępniane są na nim informacje na temat liczby, pochodzenia, zależności od czasu i głównych regionów zapytań kierowanych do wyszukiwarki Google. Przy pomocy tego serwisu można określić częstości, z jakimi są wyszukiwane określone słowa kluczowe.

Cel pracy. Celem badania było określenie możliwości przewidywania zapadalności na grypę na podstawie częstości wyszukiwań określonych słów w wyszukiwarce Google.

Materiał i metody. Przy pomocy serwisu GT pobrano informacje o dziennych częstościach wyszukiwań dla słów „grypa", „przeziębienie" i „gorączka" w latach 2014–2016 na terenie Polski. Z meldunków epidemiologicznych NIZP–PZH uzyskano dane na temat liczby zachorowań i zapadalności na grypę. Dokonano oceny korelacji pomiędzy tymi zmiennymi (współczynnik korelacji Spearmana).

Wyniki. Znamienną statystycznie korelację potwierdzono pomiędzy średnimi dziennymi wartościami wyszukiwań (ŚDWW) wszystkich 3 słów kluczowych i zapadalnością na grypę według danych NIZP–PZH. Najsilniejszy związek stwierdzono pomiędzy ŚDWW dla słowa „przeziębienie" i zapadalnością na grypę ($\rho = 0{,}77$; $p < 0{,}05$).

Wnioski. Analiza częstości wyszukiwań w wyszukiwarce Google pozwala przewidywać trendy w zakresie zapadalności na grypę. Analiza wyszukiwań w Internecie może być uzupełnieniem tradycyjnego monitorowania chorób.

Słowa kluczowe: infodemiologia, grypa, wyszukiwania internetowe

Background

Influenza is still one of the most common viral diseases, and one of the most dangerous ones. Its supervision and prevention remains a great challenge for the medical community and for public health professionals. Influenza is a viral disease of the respiratory tract, which occurs in 5–10% of the adult population and in 20–30% of children each year.[1] The problem of influenza returns with each epidemic season, and despite the progress made in medical sciences in recent decades, its incidence and severity cannot be predicted.[1,2] The influenza pandemic that occurred in 1918 was caused by a mutation of the H1N1 influenza virus. It is estimated that this variant of the influenza virus killed from 50 to 100 million people worldwide.[3]

Epidemiological surveillance plays a key role in monitoring and responding to the threat of an influenza pandemic.[4] Many organizations around the world are responsible for the surveillance, monitoring and prevention of influenza. The World Health Organization (WHO) plays a leading role in the surveillance and monitoring of morbidity and mortality related to infectious diseases. On a global scale, influenza remains a serious threat. Yearly, about 1.8 billion people suffer from influenza, and more than 500,000 die from it.[1] In Poland, the National Institute of Public Health – National Institute of Hygiene (NIPH-NIH) is responsible for collecting, analyzing and distributing information on influenza incidence.[5,6]

Current strategies of surveillance and monitoring of influenza and influenza-like illnesses do not ensure sufficient efficiency in predicting influenza incidence in the population.[7] In most countries, surveillance and monitoring of influenza is based on data collected in medical institutions reporting cases of influenza and influenza-like illnesses on the basis of medical records.[8]

New forms of surveillance and monitoring of influenza are being developed. An innovative approach is based on the analysis of the frequencies of searches of specific keywords in web search engines, most commonly in the Google search engine. This method is an example of the strategies developed within the field of infodemiology. Infodemiology is defined as the domain in which the content searched for or published in the Internet is used to analyze health-related phenomena. The main assumption is that the data accumulated in the electronic environment (mainly in the Internet) may be treated as public health information and used to set health policies.[9] In particular, infodemiology may rely on the exploration of web content to predict health trends in the population.[9] In other words, infodemiology utilizes automatically aggregated data on the incidence of and searches for specific information performed on websites and social networking sites.[10] It seems that for some diseases, especially infections, search engine data may be a source of valuable information.[11]

The Google Trends (GT) tool is an important source of data for infodemiology. This is a free and publicly available service that permits analyses of searches performed with the Google® web search engine.[11,12] The number of searches for a given keyword is compared to the total number of searches for the selected period.[8] The results of earlier studies indicate that this application may be used for real-time detection of trends in the incidence of infectious diseases, e.g., influenza or dengue. It can also help to optimize the response to outbreaks of such diseases.[13,14]

The main aim of this study was the assessment of the usefulness of information available from the GT service to predict the actual incidence of influenza in the Polish population.

Material and methods

Data retrieval

The GT service was used to retrieve data on searches for 3 keywords (the Polish equivalents of "fever", "flu" and "cold") from the period 2014–2016. The data were extracted as coefficients of daily searches and downloaded as CSV files in batches corresponding to 6-month intervals: from January 1, 2014 to July 1, 2014; from July 2, 2014 to December 31, 2014; from January 1, 2015 to July 1, 2015; from July 2, 2015 to December 31, 2015; from January 1, 2016 to July 1, 2016; and from July 2, 2016 to December 31, 2016. The data from epidemiological reports issued by the NIPH-NIH on the incidence of influenza in Poland in the same periods were also collected.

Epidemiological reports on incidence of influenza

Influenza surveillance has been conducted by the WHO since the mid-20th century. In 1996, the European Influenza Surveillance Network (EISN) was established. The EISN connects institutions that collect epidemiological and virological data on influenza. They are responsible for providing reliable information to public health experts in EU member states so they can take appropriate actions relevant to influenza activity. The EISN centers collaborate with the Ministries of Health in countries that have joined the network. The data for the reports is obtained from health care providers participating in the SENTINEL influenza surveillance program. In Poland, 1–5% of family physicians participate in the program, which is maintained by 16 Provincial Sanitary and Epidemiological Stations (PSES) and the National Influenza Center based in the NIPH-NIH as a coordinating unit. The surveillance involves a representative sample of the Polish population, and is based on reports on influenza and suspected influenza cases from the participating physicians.[15] All the physicians involved are obliged to take samples from patients with influenza-like symptoms (according to criteria established by the European Union).[16] This definition says that the influenza is characterized by "a sudden onset of symptoms and at least one of the following four systemic symptoms: fever, headache, muscle pain, confusion, and one of 3 respiratory symptoms: cough, sore throat, shortness of breath".

The data and samples collected from patients are sent to relevant PSESs. The diagnosis of influenza is confirmed by virological laboratories. Virological and epidemiological data are then sent by the PSESs to the National Influenza Center, where weekly reports for the whole country are prepared. Every year, the Center issues 48 reports including data stratified according to voivodeships and age groups.

In this paper, we used data from 144 epidemiological reports issued in the period from January 1, 2014, to December 31, 2016.

Google Search Frequencies

Using GT, one can obtain information about the relative frequencies of searches for specified keywords within certain time periods performed with the Google search engine in chosen geographical areas. The queries can be narrowed down to a specific time interval and language. The coefficients of search frequencies are presented as a time series. Raw numbers of searchers occurring in time units within the analyzed period are compared to the total number of performed searches. The resulting values are then expressed as a proportion of the maximum coefficient of search frequency in the period and displayed on a scale from 0 to 100.[12,17]

Sets of daily search coefficients for the Polish equivalents of "flu", "cold" and "fever" were generated for the period from January 1, 2014 to December 31, 2016. Then, for each period corresponding to the 144 NIZP-PZH epidemiological reports used for this study, the average daily search coefficients (ADSC) were calculated on the basis of the daily search coefficients retrieved from GT.

Statistical methods

STATISTICA PL v. 13.1 software (StatSoft, Tulsa, USA) was used for the analysis. For quantitative variables, descriptive statistics including mean and standard deviation (SD) were calculated. The distributions of the quantitative variables were checked with the Shapiro-Wilk test. Due to the non-conformance to normal distribution of the variables based on data coming from GT, Spearman's rank-order correlation was used.

Results

Trends in influenza occurrence and search frequencies

The mean ADSC (±SD) for "flu" was 33.4 ±21.1. The lowest ADSC in the analyzed period was 3.4 and the highest was 84.8. The mean ADSC for "cold" was 43.7 ±19.8 with a range from 10.4 to 80.3. For the term "fever", the mean ADSC was 40.6 ±9.4, with a range from 12.4 to 64.6. The average number of new cases of influenza reported by the National Influenza Center in the period from 2014 to 2016 was 77 820.8 ±43 318.1. The lowest reported number of cases was 17 407 and the highest was 212 660. The mean incidence of influenza in the period analyzed was 26.6 ±15.2, with a range from 5.46 to 79.0.

The peaks of influenza incidence usually occurred in mid-February (Fig. 1). The incidence then diminished until September, to slightly increase in mid-November. The peaks of searches for the term "flu" overlapped with the peaks of influenza incidence. The number of searches then dropped, descending to the lowest values in June.

From mid-September the number of searches started to rise. Interestingly, usually just before the search peak in February (overlapping with the peak of influenza incidence), a deep 2-week decrease in searches could be observed.

The trend of searches for the term "cold" tended to demonstrate 2 peaks yearly, the 1st occurring in mid-February and the 2nd in the latter half of September (Fig. 2). Between these peaks, a relatively high number of searches was maintained, without deep dives like in the case of the term "flu".

Assessment of correlations

The strongest correlation was found between the ADSC for the term "cold" and influenza incidence ($\rho = 0.73$, $p < 0.001$) and the number of new cases in reporting periods ($\rho = 0.73$, $p < 0.001$) (Table 1). Correlations between the ADSC for the term "flu" and influenza incidence, as well as between the ADSC for "flu" and the number of new cases per reporting period, were moderate ($\rho = 0.54$, $p < 0.001$ and $\rho = 0.53$, $p < 0.001$, respectively).

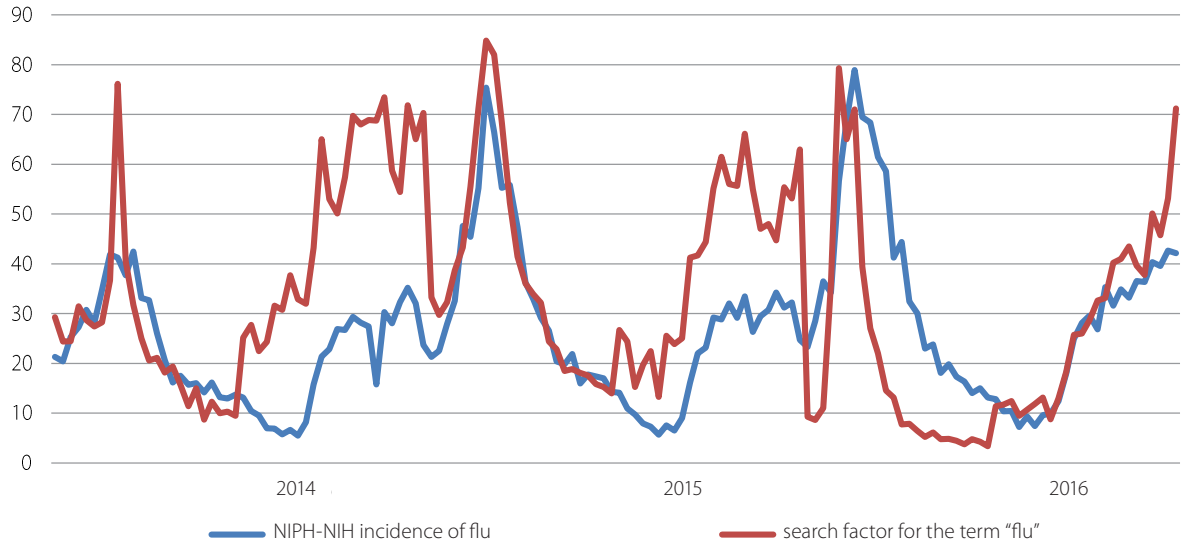

Fig. 1. Influenza incidence and the ADSC for the term "flu" in 2014–2016

Ryc. 1. Trendy zapadalności na grypę i ŚDWW dla terminu „grypa" w latach 2014–2016

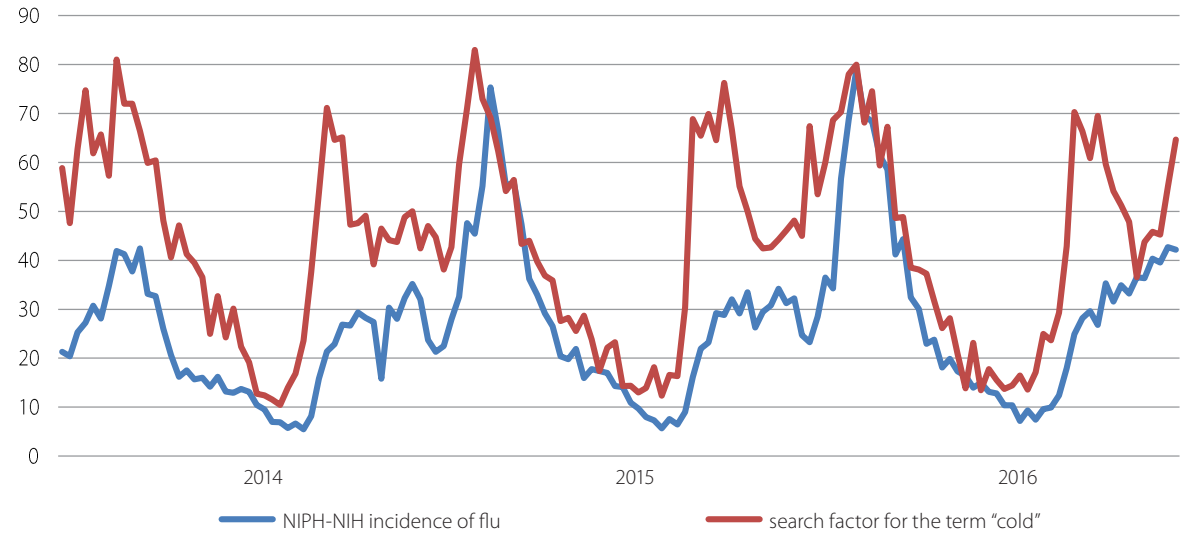

Fig. 2. Influenza incidence and the ADSC for the term "cold" in 2014–2016

Ryc. 2. Trendy zapadalności na grypę i ŚDWW dla terminu „przeziębienie" w latach 2014–2016

Table 1. Nonparametric correlation between influenza incidence and Google engine searches – ADSC

Tabela 1. Korelacja pomiędzy zapadalnością na grypę a wyszukiwaniem w Google – ŚDWW

ADSC for search terms	Influenza incidence*	Number of new cases in reporting periods*
"cold"	0.73**	0.73**
"flu"	0.54**	0.53**
"fever"	0.21**	0.22**

* According to reports issued by NIPH-NIH.
** $p < 0.001$.

Discussion

The number of studies using GT and other Internet user-generated data has increased significantly over the last few years.[11] Internet search engines have been successfully used to monitor influenza, Zika virus infections, Lyme disease, and dengue in many countries.[18–21]

Our paper demonstrated that there is a statistically significant correlation between search frequencies for selected terms and influenza activity in the Polish population. Interestingly, the strongest relationship ($\rho = 0.73$) between the search frequency coefficient and influenza incidence was found for the term "cold" and not the term "flu" as could be expected. The correlation between the search frequency coefficient for the term "flu" and influenza incidence was moderate, but still statistically significant.

Cho et al.[22] investigated the correlation between the data available from South Korea's national influenza surveillance system and the data retrieved from GT for that country. Epidemiological data for influenza originated from the Korea Centers for Disease Control and Prevention (KCDC). The authors reported a moderate and statistically significant correlation between the data retrieved from GT and the data issued by the KCDC (Pearson coefficient 0.53, $p < 0.05$).

Dugas et al.[13] developed a real-time influenza prediction model based on data from 7 years (2004–2011) and relevant data extracted from GT. The model can predict how many people will be affected by influenza a week in advance. In addition to the GT data, meteorological data and temporal information were used to develop the model.

Kang et al.[8] collected data on influenza surveillance in China from 2008 to 2011. Data on Internet searches was downloaded from GT. The authors found the strongest correlation between searches for "fever" and influenza incidence (Pearson's correlation coefficient 0.73, $p < 0.05$).

Additionally, there have been a number of studies on the relationship between traffic on other social media portals, e.g., Twitter, and influenza activity. For example, Signorini et al.[23] found that Twitter traffic can be used to track users' influenza-related concerns and to assess the activity of the disease in real time up to 2 weeks faster than data published by the EISN.

There is also growing evidence that data generated by Internet users may be useful for monitoring the activity of other infectious diseases. Chang et al.[24] found that Google searches for dengue-related terms could be used to adequately estimate the actual activity of the disease reported by institutions responsible for epidemiological surveillance in Bolivia, Brazil, India, Indonesia, and Singapore, as well as by WHO. A study by Yang et al.[14] confirmed that combining historical dengue data with searches for dengue in the Google search engine could enable medical professionals to predict the activity of the disease. They created a model that can lead to better estimates of dengue activity in real time in a self-regulating manner.

There are also some reports suggesting that infodemiology may be used for assessing the effectiveness of public health campaigns. Glynn et al.[25] found that breast cancer media campaigns generate increased numbers of searches for cancer-related keywords. Murray et al.[26] reached similar conclusions. To find out if the launch of Mouth Cancer Awareness Day in Ireland stimulated public interest in the disease, they used GT to assess the frequency of searches for "oral cancer" and "mouth cancer" in the period between January 2005 and December 2013. They confirmed that the number of searches for these phrases in the Internet increased significantly ($p < 0.001$) after launching the campaign.

Our study suffered from several limitations. First, it is an initial analysis performed for the Polish population and its scope was narrowed to 3 arbitrarily selected keywords (the Polish equivalents of "flu", "fever" and "cold"). It could be interesting to explore the relationship between epidemiological data on influenza and other terms corresponding with symptoms of the disease. Future research efforts should cover longer time intervals to check if the correlations confirmed in this paper are valid in a longer perspective. Some limitations are related to the way the data are generated from GT. Basically, daily search coefficients are available only for time intervals limited to 6 months. The comparability of search frequency coefficients generated for 6-month periods may be lower than expected. Furthermore, the data obtained from GT contains relative coefficients expressing the proportion of search frequencies for a given term in a time unit (day, week) to maximum search frequency in the retrieved period, e.g., month or year. The final values are expressed on a scale of 0–100 and each retrieved 6-month period is subject to separate indexing. In a longer perspective, the number of Internet users in a specific location or country should be also considered. When interpreting the results of our study, one should also remember that the intensity of searches for specific terms, including the context of influenza incidence, may be influenced by communication delivered by the mass media and electronic media themselves.

Despite the limitations of the study, the significant correlation between the GT data and the traditional epidemiological reports issued by NIPH-NIH indicates that the Internet may offer attractive tools for the surveillance of influenza and other infectious diseases.

Conclusions

Seasonal flu remains a challenge for public health. Globalization has led to searches for new approaches to detecting, tracking and reporting seasonal influenza and other infectious diseases. Online surveillance systems using Internet-based tools such as GT are emerging as valid disease-monitoring strategies.

The purpose of the study was to determine the feasibility of using data generated in the Internet to monitor influenza incidence in the Polish population. Data retrieved from GT has been correlated with data published by the NIPH-NIH. It seems that assessments of search frequencies for relevant terms performed using the Google search engine may be used to assess the actual incidence of influenza. It should also be noted that users of the Google search platform are prone to apply other related words, not necessarily the term "flu", in cases of influenza-like infections. This is clearly related to the overlapping clinical characteristics of influenza and the common cold. Surveillance of the frequency of flu-related searches can be a cost-effective option to complement traditional surveillance systems for this disease. Future research should adjust for the limitations signaled in this study, and for parallel phenomena that may exert an influence on search frequency, e.g., media coverage of disease activity and general penetration of the Internet in the population.

References

1. Ernst & Young. Grypa i jej koszty w Polsce. http://adst.mp.pl/s/www/opzg/Grypa-i-jej-koszty-w-Polsce.pdf. Accessed July 6, 2017.
2. Brydak LB. Grypa znana od stuleci. *Fam Med Primary Care Rev*. 2014;16(2):181–184.
3. Niall P, Johnson AS, Mueller J. Updating the accounts: Global mortality of the 1918–1920 "Spanish" influenza epidemic. *Bull Hist Med*. 2002;76(1):105–115.
4. WHO. Influenza (Seasonal). http://www.who.int/mediacentre/factsheets/fs211/en/. Accesed July 20, 2017.
5. Bednarska K, Hallmann-Szelińska E, Kondratiuk K, Rabczenko D, Brydak L. Novelties in influenza surveillance in Poland. *Probl Hig Epidemiol*. 2016;97(2):101–105.
6. NIZP-PZH. Statut Narodowego Instytutu Zdrowia Publicznego – Państwowego Zakładu Higieny. http://bip.pzh.gov.pl/public/?id=135929. Accessed July 18, 2017.
7. Moniz L, Buczak AL, Baugher B, Guven E, Chretien JP. Predicting influenza with dynamical methods. *BMC Med Inform Decis Mak*. 2016;16(1):134.
8. Kang M, Zhong H, He J, Rutherford S, Yang F. Using Google Trends for influenza surveillance in South China. *PLoS ONE*. 2013;8(1):e55205.
9. Eysenbach G. Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *J Med Internet Res*. 2009;11(1):e11.
10. Eysenbach G. Infodemiology and infoveillance tracking online health information and cyber behavior for public health. *Am J Prev Med*. 2011;40(5 Suppl 2):S154–158.
11. Nuti SV, Wayda B, Ranasinghe I, et al. The use of Google Trends in health care research: A systematic review. *PLoS One*. 2014;9(10):e109583.
12. Carneiro HA, Mylonakis E. Google Trends: A web-based tool for real-time surveillance of disease outbreaks. *Clin Infect Dis*. 2009;49(10):1557–1564.
13. Dugas AF, Jalalpour M, Gel Y, et al. Influenza forecasting with Google Flu Trends. *PLoS One*. 2013;8(2):e56176.
14. Yang S, Kou SC, Lu F, Brownstein JS, Brooke N, Santillana M. Advances in using Internet searches to track dengue. *PLoS Comput Biol*. 2017;13(7):e1005607.
15. Woźniak-Kosek A, Brydak LB. System nadzoru wirusologicznego i epidemiologicznego nad grypą w populacji polskiej – SENTINEL. *Fam Med Prim Care Rev*. 2013;15(3):483–485.
16. Commission Implementing Decision of 8 August 2012 amending Decision 2002/253/EC laying down case definitions for reporting communicable diseases to the Community network under Decision No 2119/98/EC of the European Parliament and of the Council.
17. Google Trends help. https://support.google.com/trends/answer/4365533?hl=pl&ref_topic=13975&visit_id=1-636358970435693241-2516253346&rd=1. Accessed July 18, 2017.
18. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*. 2009;457(7232):1012–1014.
19. Seifter A, Schwarzwalder A, Geis K, Aucott J. The utility of "Google Trends" for epidemiological research: Lyme disease as an example. *Geospatial Health*. 2010;4(2):135–137.
20. Ortiz JR, Zhou H, Shay DK, Neuzil KM, Fowlkes AL, Goss CH. Monitoring influenza activity in the United States: A comparison of traditional surveillance systems with Google Flu Trends. *PLoS One*. 2011;6(4):e18687.
21. Teng Y, Bi D, Xie G, et al. Dynamic forecasting of Zika epidemics using Google Trends. *PLoS One*. 2017;12(1):e0165085.
22. Cho S, Sohn CH, Jo MW, et al. Correlation between National Influenza Surveillance Data and Google Trends in South Korea. *PLoS Comput Biol*. 2013;8(12):e81422.
23. Signorini A, Segre AM, Polgreen PM. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PLoS Comput Biol*. 2011;6(5):e19467.
24. Chan EH, Sahai V, Conrad C, Brownstein JS. Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. *PLoS Negl Trop Dis*. 2011;5(5):e1206.
25. Glynn RW, Kelly JC, Coffey N, Sweeney KJ, Kerin MJ. The effect of breast cancer awareness month on internet search activity: A comparison with awareness campaigns for lung and prostate cancer. *BMC Cancer*. 2011;11:442.
26. Murray G, O'Rourke C, Hogan J, Fenton JE. Detecting internet search activity for mouth cancer in Ireland. *Brit J Oral Max Surg*. 2016;54(2):163–165.